How to (Re)represent it?

Daniel Raggi
University of Cambridge, UK
daniel.raggi@cl.cam.ac.uk

Gem Stapleton
University of Cambridge, UK
ges55@cam.ac.uk

Aaron Stockdill
University of Cambridge, UK
aaron.stockdill@cl.cam.ac.uk

Mateja Jamnik
University of Cambridge, UK
mateja.jamnik@cl.cam.ac.uk

Grecia Garcia Garcia
University of Sussex, UK
g.garcia-garcia@sussex.ac.uk

Peter C.-H. Cheng
University of Sussex, UK
p.c.h.cheng@sussex.ac.uk

***Abstract*—Choosing an *effective* representation is fundamental to the ability of the representation's user to exploit it for the intended purpose. The major contribution of this paper is to provide a novel, flexible framework, rep2rep, that can be used by AI systems to recommend effective representations. What makes an effective representation is determined by whether it expresses the necessary information, supports the execution of tasks, and reflects the user's cognitive abilities. In general, there is no single 'most effective' representation for every problem and every user, which makes it difficult to choose one from the plethora of possible representations. To address this, rep2rep includes: a domain-independent language for describing representations, algorithms that compute measures of *informational suitability* and *overall cognitive cost*, and uses these measures to recommend representations. We demonstrate the application of rep2rep in the probability domain. Importantly, our framework provides the foundations for personalised interaction with AI systems in the context of representation choice.**

I. Introduction

The ability of AI systems to adapt to human users is of major importance. This paper focuses on AI systems that recommend representations of knowledge that are *suitable* and *effective* for their intended use and the target user. The suitability of a representation is determined by the information it is meant to represent and the goals for which it is intended to be used. Effectiveness depends on the cognitive abilities of the target user. Evidentially, the development of such AI systems is challenging, necessarily requiring advances in computer science and cognitive science.

As a simple example, suppose we want to represent the following information about students: everyone who studies algebra (A) also studies topology (T) and nobody studies both topology and French (F). There are many notations for representing this information, and subsequently reasoning about it, including natural language (NL) as just stated. Others include description logic, e.g. $A \sqsubseteq T$, $T \sqcap F \sqsubseteq \bot$, and first order logic, e.g. $\forall x(A(x) \Rightarrow T(x)) \wedge \neg\exists x(T(x) \wedge F(x))$, alongside Venn, Euler and linear diagrams[1], see Fig. 1. Whilst NL is likely to be fairly effective from the perspective of understanding the information, it may not be the most effective in the context of problem solving. For instance, the Euler diagram allows one to observe that the answer to the question "does anyone who studies algebra also study French?" is no. Later in the paper (section III), we give example representations from the probability domain, complementing these which are derived from logic. Our goal is to provide the necessary foundations that support the development of an AI system that can recommend effective representations, in the context of problem solving, tailored to individual users.

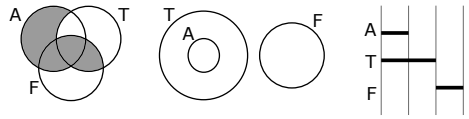


Fig. 1. Venn (left), Euler (middle) and linear diagrams.

An AI system that recommends representations needs a general theoretical framework capable of: describing representations, capturing representation-manipulation rules needed when solving problems, and understanding correspondences between representations in order to identify suitable alternatives. Beyond this, the framework must identify whether alternative choices are effective: what is effective for one user need not be effective for another. The provision of such a framework is well beyond the current state-of-the-art. Moreover, the goal of providing a complete theory of representation choice based on the cognitive abilities of users requires substantial research advances that are beyond the scope of a single paper.

The major contribution of this paper is a novel framework, called rep2rep, that can be used by AI systems to recommend effective representations. §II covers background and motivation. §III exemplifies representation choice. The core of rep2rep is in §IV: a theory that enables the description of representations. §V demonstrates how rep2rep incorporates measures of *informational suitability* and *user-specific cognitive costs*. The complete rep2rep framework (§IV and §V) encapsulates a process by which an AI system can recommend representations that are suitable for the problem and the user. The potential for the successful deployment of such a system is demonstrated in §VI: we empirically evaluate representation recommendations made by rep2rep. We conclude and present future directions in §VII.

II. Background and Motivation

In rep2rep we consider the need for a common language to talk about representations, the requirement to identify

Supported by EPSRC grants EP/T019603/1 and EP/T019034/1.

[1]In Venn diagrams, shading represents the emptiness of a set. Euler diagrams exploit spatial relations to indicate subset and disjointness. Similarly, the horizontal spatial relations between lines in linear diagrams represent subset and disjointness.

correspondences between representations, and the effectiveness of representation choice on task performance.

Representations and Problem Solving. Representations play a fundamental role in our lives. Our focus is on representations that communicate information and support users with problem solving. One basic goal is *observing* information from a representation [23, 24]. More complex problems may require representation manipulations to make deductions using (formal or informal) inference steps. To allow an AI system to compare representations, we propose a language – *representational systems* and their *descriptions* – that allows us to characterise representations and their associated manipulation rules.

Existing Frameworks. Perhaps the most prominent frameworks that support representation choice are Cognitive Dimensions [3, 7] and the "Physics" of Notations [16], but see also [4, 6]. There is *subjectivity* in how they are applied and, whilst useful, they are not able to predict the relative cognitive costs of representations and have not led to implementable AI systems. rep2rep removes the need for subjectivity through the use of measures of *informational suitability*, and takes the state-of-the-art to a new level of precision: rep2rep includes correspondences between representational systems which provide vital foundations for *objectively* comparing competing choices.

Empirical Insights. Representation choice is critical: cognitive science has established that effective representations can yield significant improvements in human task performance and learning [12, 14]. Researchers have empirically compared *specific representation choices* by manipulating graphical or topological features, as in [21]. This allows 'fine-tuning' of representations to improve their effectiveness. Moreover, *representational system choices* have been empirically compared on a limited range of representations, e.g. [15, 22, 28]. Such evaluations do not provide general insights about how to choose cognitively effective representations. This necessitates the need to improve representation choice through extensive empirical studies involving human participants. A surprising omission, which is addressed by rep2rep, is a general method that supports the objective selection of representations using measures of cognitive costs informed by the intended user and the problem they are trying to solve.

Implemented Systems. Whilst many systems support formal reasoning [10, 18], few have been implemented that aid representation choice in the context of problem solving. Reasoning tools that support multiple notations, such as [2, 26], reflect the value of alternative representational systems, but none of them specifically guide the user towards effective representation choices. rep2rep addresses this problem.

Contribution. AI systems that support effective representation choice in order to help humans solve problems are needed: rep2rep is the basis of such a system. It includes a formal conceptualisation of representations and problems, characterisation of analogical correspondences between representations, and computable measures of informational suitability and user-specific cognitive costs. This allows the objective recommendation of effective representations. Importantly, as more empirical research is undertaken, it will be possible to adapt measures of cognitive costs so that rep2rep makes ever more robust recommendations. The generality and extensibility of rep2rep ensures long-term relevance.

III. The Diversity of Representations

Representations are built from tokens that satisfy some syntactic constraints. Tokens include the numerals 3 and 12, and the symbols $+$ and $=$. Representations include $1.2+3.6=4.8$ and $1.2+3.6=7$ and they satisfy the rules that $+$ and $=$ are binary operators that join together 'valid' representations; $1.+$ is not valid. The fact that we can build a variety of representations from a set of tokens is the basis on which we will define representational systems (RSs): RSs allow us to abstractly characterise classes of representations.

To exemplify representation choices, we present a problem and different ways in which it can be expressed and solved. The scope of these representations, and thus RSs from which they are drawn, is large, including natural language (NL), formal notation, geometric figures, and tables. They have different syntaxes and manipulation rules so it is likely that the cognitive effort demanded of the user varies across representations.

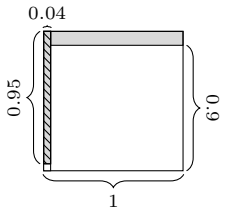
Medical Test Problem (represented using the NL RS) 4% of the population has a disease D. For those who have the disease, a test T is accurate 95% of the time. For those who do not have the disease, T is accurate 90% of the time. If you take the test and it comes out positive, what is the probability that you have the disease?

Alternative Rep. 1 (represented using the Bayesian RS)**.** *Assume:* $\Pr(D)=0.04,\ \Pr(T\mid D)=0.95,\ \Pr(\overline{T}\mid\overline{D})=0.9$. *Calculate:* $\Pr(D\,|\,T)$.

Solution.

$$\begin{aligned}\Pr(D\mid T)&=\frac{\Pr(T\mid D)\Pr(D)}{\Pr(T)}=\frac{\Pr(T\mid D)\Pr(D)}{\Pr(T\cap D)+\Pr(T\cap\overline{D})}\\&=\frac{\Pr(T\mid D)\Pr(D)}{\Pr(T\mid D)\Pr(D)+\Pr(T\mid\overline{D})\Pr(\overline{D})}\\&=\frac{\Pr(T\mid D)\Pr(D)}{\Pr(T\mid D)\Pr(D)+(1-\Pr(\overline{T}\mid\overline{D}))(1-\Pr(D))}\\&=\frac{0.95\cdot0.04}{0.95\cdot0.04+(1-0.9)\cdot(1-0.04)}\approx0.28\end{aligned}$$

Alternative Rep. 2 (represented using the Areas RS)**.** *Calculate the size ratio between the patterned region,* ▨*, and the shaded region,* □*.*



Solution. From the figure, observe that the area of the patterned rectangle is 0.95×0.04. The area of the two shaded rectangles, (including the patterned region), is $0.95\times0.04+(1-0.9)\times(1-0.04)$. The ratio of these areas is ≈0.28.

Alternative Rep. 3 (represented using the Contingency Table RS)**.** *Calculate the ratio of the value of cell* (T,D) *against* $\mathrm{total}(T)$ *(see the shaded cells).*

	D	$\overline{D}$	**total**
T	$0.95\cdot0.04$		
$\overline{T}$		$0.9\cdot\mathrm{total}(\overline{D})$	
total	0.04		1

Solution. Using the law of total probability, calculate $\text{total}(\overline{D})$, followed by the value of cell $(T, \overline{D})$, followed by $\text{total}(T)$. The desired ratio is ≈ 0.28.

These representations exemplify the problem our framework aims to solve. Suppose an AI system is provided with the NL representation. How do we know which alternative RSs may be *informationally suitable*: which ones can express the required information and deliver a solution using its laws and tactics? In fact, all of the examples are suitable, but many other RSs exist which are not. rep2rep, when provided with information about correspondences between RSs and cognitive costs, can identify suitable and cognitively effective alternatives.

IV. The rep2rep Framework

Fig. 2 illustrates the architecture of the rep2rep system. An analyst will be required to exploit so-called *representational system (RS) descriptions*, each of which describes the *set of representations* that can be formed over some vocabulary given some syntactic construction rules. It also describes rules than can be used to manipulate representations when solving problems. Using an RS-description, the analyst can provide a description of a *specific* problem; this is called a *Q-description* in Fig. 2. The rep2rep framework takes the Q-description and RS-descriptions of alternative RSs and computes measures of information suitability and overall cognitive cost that can be used to rank RSs, thus providing a recommendation to the analyst. In addition, these rankings are user-specific, in that they account for different user profiles.

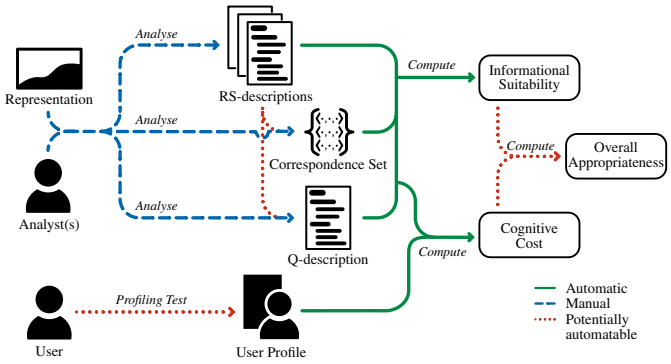


Fig. 2. Architecture of the rep2rep AI system.

One particular RS is that of Bayesian probability. This RS comprises a set, say $\mathcal{E}$, of basic events, all of which are representations within this RS. Basic events can be combined to form more complex events, using the binary operators, $\cap$ and $\cup$, alongside the complement operator $\overline{\ }$. Given the resulting (inductively defined) set of events, formed using these operators and brackets, to avoid ambiguity, further representations can be built. Given events E_1 and E_2, further representations include: $\Pr(E_1)$, $\Pr(E_1|E_2)$, $\Pr(E_1) = \Pr(E_2)$, and $\Pr(E_1|E_2) = 0.5$. Other arithmetic operators are also included such as $+$ and $-$. The set of representations within the Bayesian RS includes (a) all events, (b) the Bayesian and arithmetic operators, (c) real numbers, and (d) all of the representations thus formed, following the obvious inductive construction rules.

rep2rep aims to *describe*, not formalise, representations and RSs with *enough* detail to enable effective representation selection. Formalisation is a costly process, and rep2rep is intended for flexible use where non-formalised representations, such as a diagram in a textbook, can be described. At the core of rep2rep are the notions of RSs, which characterise classes of representations, RS-descriptions, which encode important features of RSs, and Q-descriptions[2] that abstractly encode the representation of a problem that is being solved. The requirements for RSs, RS- and Q-descriptions, and an informal presentation of them, were first given in [19, 20]. We extend that work by precisely specifying RSs, Q- and RS-descriptions.

A. Abstract Characterisation of Representations

The representations characterised by an RS are formed over a common syntax, which may be defined by grammatical rules, along with rules of inference. We specify an RS by its *components*: *terms* (*primitive* and *composite*), *valid expressions*, *types*, *laws*, and *tactics*[3].

Primitive terms, also called *primitives*, are elemental pieces from which representations are built (e.g., $\Pr$, $|$, D, $=$, 0.95). From primitives, composite terms are constructed (e.g., $\Pr(T \mid D)$). Valid expressions, or simply *expressions*, are terms that abstractly characterise representations (i.e., the concrete instances of an RS, such as $\Pr(T \mid D) = 0.95$). Types classify terms into categories which allow the RS's grammar to be specified. For instance, replacing 0.95 by $+$ in $\Pr(T \mid D) = 0.95$, we obtain a construction which is not an expression (or a term) within the Bayesian RS. The type of 0.95 is `real`[4], and anything that takes its place must have the same type.

Laws allow inferences to be made. In the Bayesian RS, the equality law states that if two expressions are equal then one may be replaced by the other in any expression. Another law states that $\Pr(x \cap y) = \Pr(x \mid y)\Pr(y)$. Tactics are tools for manipulating expressions, and laws are units of knowledge that enable their use. For example, the tactic *rewrite* allows us to replace the term $\Pr(T \cap D)$ with $\Pr(T \mid D)\Pr(D)$, due to the aforementioned laws.

B. RS- and Q-Descriptions

To facilitate the implementation of rep2rep in AI tools, we computationally model the abstract concepts of representations and RSs with *RS-descriptions* which describe complete RSs, and *Q-descriptions* which describe specific representations of problems. Both kinds of descriptions are comprised of a set of *declarations*. A declaration is a triple, (k, v, A), where k is one of the following kinds: *primitive*, *type*, *law*, *tactic*, or *pattern*; v is a specific value of k; and A is a finite set of *attributes*, written as $\{a_1 := x_1, \ldots, a_n := x_n\}$. Here, we adopt a more

[2] 'Q' for *question*; [19] refers to 'descriptions' as 'tables'.

[3] Prior work called primitive terms 'tokens'; primitive terms are an abstraction of a representation's tokens. Composite terms and expressions were just called *expressions*.

[4] Or more accurately, it is a base-10 real numeral.

readable notation: $k\ v : \{a_1 := x_1, \ldots, a_n := x_n\}$. In the Bayesian RS, we can declare

$$\begin{aligned} \text{primitive } 0.95 &: \{\text{type} := \texttt{real}\} \\ \text{primitive } D &: \{\text{type} := \texttt{event},\ \text{occurrences} := 4\}, \end{aligned}$$

with the latter being for the medical problem's representation. Intuitively, attributes provide structural information or identify features. In an RS-description, the attributes of a kind, k, and its value, v, assert general information about v. In a Q-description, the attributes also assert representation- and problem-specific information, such as how many times v occurs. Thus, the declarations of a Q-description are taken from the associated RS-description, but include more attributes.

Pattern declarations are used to compactly describe sets of terms that have some common structure and they involve three attributes: *type*, *holes*, and *primitives*. We explain their roles by example: consider the *conditional probabilities* structure: $\Pr(\_|\_) = \_$. For this structure, we declare a pattern with name `CP` and its associated expressions are all of type `formula`. There are three *holes* that must be filled: two with terms of type `event` and one with a term of type `real`. Conditional probability expressions all use the following primitives: $\Pr, |, =, (,)$. Thus, we declare pattern `CP` as

$$\begin{aligned} \text{pattern } \texttt{CP} : \{&\text{type} := \texttt{formula}, \\ &\text{holes} := [\texttt{event}, \texttt{event}, \texttt{real}], \\ &\text{primitives} := [\Pr, |, =, (,)]\}. \end{aligned}$$

Note that a pattern abstracts away the order in which primitives and holes appear. This is by design as patterns are meant for general purpose use. For instance, in the Areas RS, we can capture an emergent new region by taking the union of two existing regions:

$$\begin{aligned} \text{pattern } \texttt{region\_union} : \{&\text{type} := \texttt{region}, \\ &\text{holes} := [\texttt{region}, \texttt{region}], \\ &\text{primitives} := []\}. \end{aligned}$$

Suppose we have an RS (e.g., Bayesian RS), with an associated RS-description, r, that we would like to use for representing and solving a problem (e.g., Medical test). Some parts of this RS, and associated laws and tactics declared within r, are necessary or helpful for this specific problem, whereas others are irrelevant. The rep2rep framework captures this in Q-descriptions by using representation- and problem-specific measures of *informational importance*: a Q-description for a representation and problem is a set of declarations, $\mathscr{D}_{q_r}$, together with a problem-specific function,

$$\text{importance}_{q_r} : \mathscr{D}_{q_r} \to [0, 1],$$

which indicates the importance of the concept captured by each declaration, relative to the problem to be solved.

For instance, in the Bayesian representation of the medical problem, the primitive $\Pr$ is of high importance because every probability problem depends fundamentally on the definition of the probability function $\Pr$. By contrast, the primitive 0.04 is of lower importance as the specific number may change the end result, but does not change the nature of the solution. The law *Bayes' theorem* has high importance as it is needed to solve the problem. By contrast, the *law of monotonicity* (if $A \subseteq B$ then $\Pr(A) \leq \Pr(B)$) is not needed here, so has low importance.

The values for informational importance must currently be set by an *expert analyst* who wants to deploy the rep2rep framework. Future work will explore whether and how these values can be inferred automatically.

V. Informational Suitability and Cognitive Costs

We now demonstrate how rep2rep makes recommendations: (i) given a Q-description of a problem in some representation, and RS-descriptions of alternative RSs, *estimate the relative informational suitability* of each of these RSs; and (ii) given a set of Q-descriptions (the problem encoded in alternative representations), *estimate their relative cognitive costs*.

A. Informational suitability

Suppose we have a problem (e.g., Medical test), represented in some RS (e.g., Bayesian). Moreover, suppose we have a set of alternative RSs (e.g., Areas, NL), but we do not know whether they can *represent* our problem. Knowing the relationship between the original RS and the alternative RSs is crucial for assessing whether they can represent the problem. For example, if there exists a formal translation of representations in the original RS into an alternative RS then the problem can be translated and, thus represented. If, in the alternative RS, we can also find laws and tactics for solving the problem then it is *informationally suitable*. Full knowledge about the expressiveness of RSs and the problems solvable using their laws and tactics is rarely available – especially when some systems have not been fully formalised, but are merely described. Thus, rep2rep exploits RS- and Q-descriptions for identifying relationships, using the novel notion of *correspondences* [19, 25] between RSs, in the context of the given problem to be solved.

As well as using the declarations in RS- and Q-descriptions, correspondences also use composite declarations (from this point, just called declarations), formed using the operators AND, OR and NOT. For example, we can write a declaration (for simplicity we omit attributes) from the RS-description of NL: '(primitive disease) OR (primitive illness)' and we say that it is satisfied by a Q-description, q_r, if one or both of the primitives 'disease' and 'illness' are declared in q_r. Correspondences between declarations from different RSs capture the analogical mappings between RSs. They are accompanied by a parameter *strength*, s_i where $s_i \in [0, 1]$, to quantify the strength of this analogical mapping. Here are some examples:

- primitive: *intersection* and *and* in NL are analogous to $\cap$ in the Bayesian RS; the correspondence is written as: $\langle \text{primitive intersection OR primitive and}, \text{primitive } \cap, s_1 \rangle$;
- types: every *event*, in the Bayesian RS, can be represented by a *region* in the Areas RS: $\langle \text{type event}, \text{type region}, s_2 \rangle$;
- laws: the additivity of disjoint probabilities (Bayesian RS) corresponds to the additivity of the areas of disjoint regions (Areas RS): $\langle \text{law Pr\_additivity}, \text{law area\_additivity}, s_4 \rangle$.

Here, s_1, s_2, s_3 and s_4 are likely to be high (close to 1) since the analogies are strong.

A *correspondence* between two RS-descriptions, r and r_i, is a triple, $\langle \alpha, \beta, s \rangle$, where α and β are declarations stemming from RS-descriptions r and r_i, respectively, and $s \in [0, 1]$. Generally, strengths may be informed by theoretical or empirical findings; for a probabilistic computation method, its statistical interpretation and some of its provable consequences, such as reversibility, composability, and extendability, see [25].

Consider a problem with Q-description q_r and some alternative RS described by r_i. Each correspondence $\langle \alpha, \beta, s \rangle$ – where α is satisfied by q_r and β stems from r_i– indicates the suitability of the alternative RS to represent (using β) the aspect of q_r captured by α. Combining correspondences to measure the extent to which informationally important aspects of the problem can be represented in the alternative RS is challenging.

To measure informational suitability, the work in [19] combined correspondences by summing their strengths multiplied by the importance of associated declarations in the Q-description. We refine this to exclude superficially different, and thus redundant, correspondences from the computation. We need a set of correspondences, C, between r and r_i, that is **minimally redundant and maximally covering** (MRMC) with respect to q_r. This is defined to be the case if the following conditions are met, where we say that the elements of $\{\alpha : \langle \alpha, \beta, s \rangle \in C\}$ are the left-hand formulae of C:

1) the left-hand formulae of C are satisfied by q_r.
2) declarations in q_r should be *maximally covered* by the left-hand formulae of C, maximising the importance and strength factors;
3) the information given by the declarations in q_r, covered by the left-hand formulae of C, should be *minimally redundant*.

Finally, we can define a **relative measure of the informational suitability (IS)** of a candidate representation with RS-description, r_i, for a given problem with Q-description, q_r, arising from RS-description r, and a MRMC set of correspondences C between r and r_i:

$$\mathrm{IS}(q_r, r_i, C) = \sum_{\langle \alpha, \beta, s \rangle \in C} \mathrm{importance}_{q_r}(\alpha) \cdot s \qquad (1)$$

In practice, finding MRMCs is computationally challenging; we currently use approximations – developing heuristics for this is left for future work.

The IS measure gives preference to RSs that can cover the important aspects of a representation of a problem, whilst also giving preference to stronger analogies. Note that IS is a *relative* measure: it is only meaningful when comparing the score for different candidate RSs with respect to the same problem. Moreover, it depends on the choice of descriptions and correspondences. Ultimately, the quality of heuristic solutions will profoundly influence the robustness of the measure.

B. Cognitive costs

The IS measure allows us to compare a set of alternative RSs described by $r_1, ..., r_n$ for re-representing a problem described by q_r. Let us assume we have Q-descriptions, $q_{r_1}, \ldots, q_{r_n}$ for the re-represented problem. These q_{r_i} could be automatically derived from correspondences and q_r (future work) or, when such automation is not possible, correspondences could serve as hints for an expert analyst, who wants to deploy the rep2rep framework, to create them.

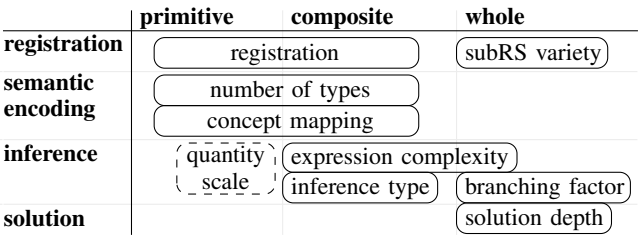


Fig. 3. Cognitive properties organised according to notation granularity (columns), and level of cognitive processing (rows).

For each r_i, we want to estimate the *overall cognitive cost* demanded of the user solving the problem[5] described by q_{r_i}. Cognitive costs arise from: (a) registering the components of a representation, (b) parsing structure, and interpreting symbols, and (c) using the laws and tactics available to derive a solution. We refer to the properties of representations giving rise to these costs as *cognitive properties*. Most cognitive properties can be determined from Q-descriptions and correspondence sets, others require problem-solutions. Solution-specific information is given by additional attributes in Q-descriptions, such as the number of *occurrences* of the tactic's use in the solution. Each property is assumed to contribute independently to the overall cost when obtaining an *overall measure of cognitive cost*; we will consider interaction among properties in future work.

We present a set of *core* measures of cognitive costs based on established cognitive phenomena[6]. They reflect two dimensions at which cognitive processes happen: *notation granularity* (spatial) and *level of cognitive processing* (temporal) [1, 17]. The *cognitive properties* of representations that we assess fall within this 2-dimensional model (see Fig. 3).

We introduce the concept of *gravity*, defined over the declarations in a Q-description, q_{r_i}. In particular, the gravity of each primitive, type and pattern is its number of occurrences in the representation (taken from the *occurrence* attribute of declarations in q_{r_i}) times its importance, as specified in q_{r_i}. The gravity of a law or a tactic is the number of times it is used in the known solution times its importance. Gravity moderates cognitive costs to be proportional to the importance and the number of occurrences of the component of the representation embodied in the declarations. The gravity of declaration x is denoted $w(x)$.

We proceed to briefly explain nine *cognitive properties* with their associated *cost functions*. The properties are about registering primitives and grouping them, interpreting compo-

[5]We assume a solution is available in the system. The challenge of estimating the overall cognitive cost of solving a problem without knowing a solution is the subject of future work.

[6]This set is not exhaustive: producing a *complete* measure of overall cognitive cost for *any* representation and problem is well beyond state-of-the-art.

nents and parsing their structure, applying tactics and laws, and domain-specific cognitive effects. The cost function for each cognitive property uses values assigned to attributes in declarations of q_{r_i} (e.g., tactic has an attribute *inference type* that can take on the value *assign*, *match*, *substitute*, *calculate*, or *transform*), and additional parameters specific to the property (e.g., the cost of using a tactic with inference type *calculate* relative to a tactic with inference type *substitute*). The values of attributes and parameters should be empirically informed for a robust calculation, but this is out of the scope of this paper (we assign provisional values informed by the literature) since our focus is on building the general framework.

Registration This is the cost of the user identifying, acknowledging and noting the location of a primitive or term in a representation [7, 13]. The registration of a primitive is characterised by the patterns in which it appears. Thus, for each primitive in the pattern, an attribute *primitive registration* assigns it a value of *icon*, *notation index* or *search*. The registration cost, $r(a)$, for primitive a is assumed to be lowest for icon and highest for search (e.g., icon = 1, notation ind. = 2, search = 4). This assignment is not intrinsic to a, but to the (potentially multiple) patterns in which a appears, Thus, to compute $r(a)$ from a Q-description, all the pattern declarations where a occurs are collected; each has a primitive registration attribute (e.g., search), which yields an individual cost (e.g., 4). The average of these individual costs–weighted by the pattern's gravity–yields $r(a)$. The registration cost for all primitives is $\sum_a w(a) \cdot r(a)$; this cost is defined analogously for terms.

subRS variety This refers to a measure of heterogeneity within a representation, in particular, how many different sub-systems need to be taken into account by the user [27]. High heterogeneity involves a heavy cost. For example, the Contingency Table RS relies on a subRS associated with arithmetic expressions, and another subRS associated with tabular organisation.

Number of types Identifying the types of terms is part of processing the semantics. A larger variety in the *types* of terms, in many cases, means a higher semantic processing cost. In rep2rep, empirically-derived knowledge about each RS is used to define the associated cost function.

Concept-mapping This property refers to the semantic mapping of terms to their corresponding concepts [28, 8, 16]. Its cost reflects the processing of various conceptual shortcomings: *deficit* (a concept with no representing symbol), *redundancy* (two symbols for one concept), *excess* (a symbol that does not map to an important concept) or *overload* (one symbol for multiple concepts). For a Q-description, q_{r_i}, we can estimate the relative number of each of these conceptual shortcomings by comparing q_{r_i} to some *reference Q-description* q' using correspondences.[7] Each conceptual shortcoming results in a penalty. The total concept-mapping cost is the gravity-weighted sum of the penalties.

[7]The reference can, in principle, be any Q-description (e.g., the original q_r). The calculations of the shortcomings are relative to the reference, so it is preferable, for an accurate calculation, that the reference has fewer concept-mapping shortcomings (if this can be known in advance).

Expression complexity This measures the complexity of the terms generated from patterns. Specifically, we measure how many nodes are in potential parse trees. Our algorithm takes each pattern and instantiates its holes recursively with type-appropriate patterns or primitives until no holes remain uninstantiated. To ensure termination, instantiation is limited by the occurrences of parts of the representation. This process results in *parse trees* for expressions. We generate, for every pattern, a sample of trees that satisfy it. The average number of nodes in the trees measures the pattern's complexity. Given a Q-description, we estimate the complexity for each pattern, and combine them into a measure of expression complexity.

Inference type This relates to the difficulty of applying tactics [1, 11, 17]. The *inference type* of tactics is captured by an attribute. Each kind of inference type is associated with a cost. A typical cost order is: *assign* (lowest), *match*, *substitute*, *calculate*, *transform* (highest). Each individual tactic gets a cost from its inference type attribute. The total cost, for a solution, is the gravity-weighted sum of the individual costs for each of the tactics. Future fine-tuning is possible to reflect the laws used within each tactic.

Branching factor This refers to the breadth of possible tactic applications in the search for a solution. A higher branching factor results in a higher cost. It is estimated from the number of tactics and the multiple ways of applying them (using different patterns or different laws).

Solution depth This is a solution-specific measure, simply calculated as the total number of tactic applications (from the tactic attribute *occurrences*).

Quantity scale The above eight properties are based on *general* insights from cognitive science. For concepts related to quantity, the scale hierarchy – *quantity*, *nominal*, *ordinal*, *interval ratio* – is known to have increasing cognitive costs [29]. We estimate these costs according to the correspondence-mapping between a Q-description and arithmetic concepts: $<, >, \leq, \geq, \max, \min$ are associated with ordinal; $+, -, \sum$ are associated with interval; and $\cdot, \times, *, \div, /, \prod, \gcd, \mathrm{lcm}$ are associated with ratio quantity scale. Further domain-specific costs can be included in the future.

C. Integrating cognitive costs

The measures of cognitive costs associated with the cognitive properties are independent (by design). So, a measure of total cost can be obtained by summing the costs, provided they are normalised in an empirically-informed way. Thus, we need a function, norm_p, that normalises the cost of each property, p, relative to the range of cognitive costs computed for the alternative representations that we are comparing. A *user-independent* measure of the total cost of a representation with Q-description q_{r_i} is given by

$$\sum_{p \in P} \mathrm{norm}_p(\mathrm{cost}_p(q_{r_i}))$$

where P is the set of cognitive properties (see Fig. 3) and cost_p is the cost function for p. Our working assumption (that needs to be fine-tuned based on empirical findings) for norm_p

is that as we move higher in notational granularity and higher in cognitive level, the cost is more substantial.

Modelling individual users We adjust the measures of cognitive costs for individual users according to their *expertise* [5]. Expertise is encoded as a number $u \in [0, 1]$; higher values mean a higher level of expertise. Expertise impacts how the *importance* values are used to compute costs. It is also related to the *granularity* dimension of the property and, thus, its cost.

Importance discernment (heuristics) A novice user may not have the expertise to discern informational importance as well as experts. Thus, we adapt the function $\mathrm{importance}$ to the user so that $\mathrm{importance}_u$ is flatter for more novice users. Specifically, we use

$$\mathrm{importance}_u(x) = 1 + \mathrm{importance}(x) \cdot u - u.$$

Thus, for the least competent users ($u = 0$) we have $\mathrm{importance}_u(x) = 1$ for any component x: every component seems equally important. For experts ($u = 1$), we have $\mathrm{importance}_u(x) = \mathrm{importance}(x)$, which is simply the informational importance as given in the Q-description: the user understands the informational importance of x.

For example, consider the property *branching factor*: if only important patterns are identified and exploited then the branching factor is reduced. This is equivalent to having good heuristics. Expert users do not explore all of the patterns, but use heuristics to identify those which are relevant to prune the search space effectively. By contrast, a novice who cannot differentiate which laws and tactics are useful, will need to explore them all when deriving a solution. Now, some cognitive cost calculations use gravity, which in turn uses importance, so flattening importance affects the cognitive costs.

Granularity-sensitive weight differentials Cognitive processes involving composite terms and expressions, rather than primitives, are more sensitive to expertise [5]. In other words, the properties to the right in Fig. 3 are more influenced by expertise than those to the left. For example, the cost of registering primitives is similar across all levels of expertise, whereas registering composite terms is more costly for novices. We model this with a multiplicative weight, $c_p(u)$, per cognitive property p: for the lowest granularity, $c_p(u)$ is equal for both experts and novices, whereas for a higher granularity, $c_p(u)$ is higher for novices than for experts.

Finally, we are able to define how to calculate the overall measure of cognitive cost of using a representation to solve a problem (encoded with Q-description q_{r_i}), given a user u, as:

$$\mathrm{Cost}(q_{r_i}, u) = \sum_{p \in P} c_p(u) \cdot \mathrm{norm}_p(\mathrm{cost}'_p(q_{r_i}, u)), \qquad (2)$$

where $\mathrm{cost}'_p(q_{r_i}, u) = \mathrm{cost}_p(q'_{r_i})$, and q'_{r_i} is the result of replacing $\mathrm{importance}$ in q_{r_i} with $\mathrm{importance}_u$.

Empirical tests are needed to ascertain the robustness of these weights and the resulting cost function. But, the take-away message from this section is that rep2rep incorporates sophisticated approaches – based on Q-descriptions and the cognitive science literature – to compute measures of cognitive costs for competing representations, measures of informational importance, and correspondences.

VI. Evaluation

rep2rep is the first to lay foundations for the computational analysis of representation choice, so cannot be compared to existing systems. Hence, we present an empirical study, comparing computed measures of informational suitability and overall cognitive cost, to data obtained from surveying expert analysts[8]. The evaluation focuses on the probability domain, and is based on the examples in section III for the medical problem.

A. Informational suitability

We computed the IS measures (Equation 1) for the NL, Bayesian, Areas, and Contingency Table RSs. This required a *Q-description* of one representation (we used a Q-description of the Bayesian representation given in section III), *RS-descriptions* of the remaining RSs, and *MRMC sets of correspondences* that link them (for this evaluation, we selected these sets manually). We set the values for the informational importance function within the Bayesian Q-description and the correspondence strengths based on our expertise. Table I shows the computed measures; the scores from surveyed expert analysts are discussed below.

B. Cognitive costs

Cost (Equation 2) is computed using RS- and Q-descriptions for each associated RS. Again, we set the parameters for these calculations based on our expertise. Notably, we had three user profiles (novice, average and expert, described below) and, to compute Equation 2, we used the following values of u for each profile: novice, $u = 1/6$; average, $u = 3/6$; expert, $u = 5/6$. Table II shows the computed measures; the scores from surveyed expert analysts are discussed below.

C. Design and method

The goal is to see whether IS and Cost, using the medical problem and associated RSs, produce similar rankings to, and are significantly correlated with, profiles obtained by surveying expert analysts. We recruited analysts (N = 11), who were affiliated with the University of Sussex's Engineering and Informatics School or the University of Cambridge's Department of Computer Science and Technology. They all confirmed that their "day-to-day work involved a lot of dealing with maths." Each analyst was either a PhD student or an academic staff member (researcher, lecturer or professor). They completed the study online using Qualtrics.

The survey contained two tasks. In task 1, participants gave feedback regarding the *informational sufficiency* of each of the RSs. Each participant was presented with the medical problem, a *textual description* of an RS[9], and a question. The description comprised four short sentences stating how the RS encodes (i) variables, (ii) relations, (iii) probability, and (iv) operations. The participants were also shown a small representative icon of

[8]These experts were not trained in how to use rep2rep, but their expertise profile makes them target analyst-users.

[9]Euler diagrams were also described, but many mistook them as a geometric representation where size is relevant, which was not intended.

	analyst scores Mean (SD)	computed measures
Bayesian	6.0 (1.3)	17.4
Areas	4.8 (1.9)	11.4
Contingency	4.9 (1.9)	8.4
NL	3.5 (2.0)	6.9

TABLE I
SURVEYED SCORES AND INFORMATIONAL SUITABILITY MEASURES.

	expert		**average**		**novice**	
	analyst Mean (SD)	comp	analyst Mean (SD)	comp	analyst Mean (SD)	comp
Bayes	1.5 (0.7)	39.4	2.4 (1.5)	51.5	3.5 (0.9)	73.0
Areas	2.8 (1.3)	77.5	2.7 (1.2)	59.9	2.3 (1.3)	35.2
Cont	3.5 (1.2)	84.0	3.3 (1.3)	106.3	3.4 (1.4)	128.0
NL	3.9 (1.4)	89.6	4.0 (1.2)	112.0	3.4 (1.9)	134

TABLE II
SURVEYED RANKINGS AND COMPUTED MEASURES OF COGNITIVE COSTS.

the representation that was used to support recognition of the RS. An example of a description, for the Contingency Table RS, is: (i) uses rows, columns and cells for events; (ii) uses cells and relative positions of cells to encode relations among events; (iii) numbers in the cells encode values of probability; and (iv) arithmetic expressions in cells give relations among probabilities. Care was taken to keep descriptions as uniform as possible across RSs in terms of the number of words (mean: 31.6 words; fewest: 26; most: 35) and complexity of sentences (mean Flesch-Kincaid Grade Level across all RS descriptions: 9.7; lowest: 8.3; highest: 11). The participants were asked a question: "To what degree is this representation sufficient for solving the problem?" A 7- point Likert scale was used to respond (1: extremely insufficient; 7: extremely sufficient). The task was repeated for each RS and the order of the presented RSs was randomised across participants.

In task 2, participants had to rank the RSs based on user profiles. The RS descriptions used in task 1 were, for consistency, also used in task 2. The order in which the RS descriptions were presented was randomised for each user-profile question. Each question was presented on its own screen and the participant was asked, in this order (paraphrased here): 1) is it adequate for explaining the solution to a *novice* (profile: secondary school maths; 14 years old); 2) is it adequate for explaining to an *expert* (profile: holder of a science or engineering degree)?; and 3) is it adequate for explaining to an *average* student (profile: post- secondary, pre-university entrance; 18 years old)?

D. Results

In order to analyse the responses from the participants, each response was taken to be a score. In particular, for task 1, the Likert scale response given by the participants was used as the score assigned to the RS (see above). For task 2, the rank order was used as the score, with the highest ranked RS scoring 1, and the lowest ranked RS scoring 5. The resulting mean scores and standard deviations (SDs) are in Table I, used to derive an informationally suitability ranking of RSs. Table II shows means from which the analysts' RS ranking for different user profiles can be derived.

We now seek to ascertain whether the computed rankings correlate with the analysts' responses. Concerning IS, the mean analysts scores were compared with computed scores for each RS (see Table I). Both approaches ranked Bayesian and NL as the most, and respectively least, suitable. However, there is disagreement over the Areas and Contingency Table RSs: the expert analysts found them similarly suitable, whereas IS ranked the Areas RS as more suitable. There was good agreement on the rank-order, but a one-tailed Pearson's correlation test gave $r = 0.89$, with $p = 0.053$: there is not a significant correlation between the mean Likert scores and the IS measures. The low significance value reflects the small number of RSs being ranked, but we can conclude that the rank-order produced by our framework is sensible. Indeed, it would be interesting if a future, larger, study could give insight into any disagreement between the experts' and rep2rep's rankings.

Concerning cognitive costs, for each user profile, the surveyed analysts' rankings were combined by taking means (see Table II): a low mean score indicates a higher ranking, that is, the associated RS was judged more effective. Likewise, a lower *computed* Cost measure suggests a more effective RS. In Table II, we observe that rep2rep's rankings are identical to those derived from the analysts' responses in the case of the expert and average profiles, but deviate in the case of novices. The correlation between the analyst and computed values for expert and average users is high, with statistical significance, while the novice correlation is lower and not significant (at 5%): for **expert**, $r = 0.97$ $(p = 0.01)$; for **average**, $r = 0.94$ $(p = 0.02)$; and for **novice**, $r = 0.76$, $(p = 0.1)$. Disagreement in rankings or lack of correlation suggests further studies are needed. One possible explanation is that users' *familiarity* with an RS is not yet modelled by rep2rep: analysts knew that novices were unlikely to be familiar with the Bayesian RS. This suggests that user profiling could include an indication of familiarity.

E. Summary

The evaluation supports the claim that an AI system based on rep2rep can recommend effective RSs. The results are promising given they are based on measures informed by our expertise. We expect stronger results when more informed measures are derived, in part based on a deeper understanding of the cognitive abilities of a range of user profiles as well as more sophisticated empirical approaches.

VII. CONCLUSION AND FUTURE WORK

This paper makes an advance that is necessary for AI systems to be able to recommend effective representations. In particular, rep2rep is the first system that can recommend representations, in the context of problem solving, that are tailored to individual users. rep2rep has a number of novel features. It includes a theoretical conceptualisation of representational systems. By exploiting RS- and Q-descriptions, rep2rep is able to identify alternative RSs based on our theory of correspondences, alongside measures of informational suitability. Significantly, rep2rep is able to recommend RSs, or even specific representations, based on user-specific cognitive profiles and the particular problem to be solved.

Demonstrating the utility of rep2rep, our empirical evaluation revealed that resulting recommendations were well-aligned with those of expert analysts. These are promising results, particularly as there is ample potential for the derivation of enhanced measures. It will be exciting, in the future, to define more robust measures of user-specific cognitive costs as well as more sophisticated models of users beyond simply their expertise.

There are significant avenues of further research. At present, analysts must generate RS- and Q-descriptions: the ambition is to do this automatically. We are actively working on the automatic derivation of correspondences and ways of measuring informational importance. In addition, there is a wide variety of application areas for rep2rep, reflecting the goal to improve AI-human interaction in the context of representation choice. This includes everyday use of AI systems (such as devices for navigation) and specialist use (such as scientific software). For example, rep2rep could be used to build AI assistants at the interface between a scientist and a theorem prover or formal ontology. Another area is education, where the teacher (taking the role of analyst) uses a multi-representational tutoring system, and user profiles would be developed for the students. This would support student-tailored representation choices to aid their individual learning. Ultimately, and importantly, the vision is that rep2rep has the potential to be exploited in any area in which representations of knowledge are used and for which alternatives may be more effective.

References

[1] J. R. Anderson. Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science*, 26 (1):85–112, 2002.

[2] D. Barker-Plummer, et al. Openproof-a flexible framework for heterogeneous reasoning. In *Diagrams*, pages 347–349. Springer, 2008.

[3] A. Blackwell, T. Green. Notational systems–the cognitive dimensions of notations framework. *HCI models, theories, and frameworks: toward an interdisciplinary science. Morgan Kaufmann*, 2003.

[4] P. C. Cheng. What constitutes an effective representation? In *Diagrams*, pages 17–31. Springer, 2016.

[5] M. T. Chi, R. Glaser, M. J. Farr. The nature of expertise. Erlbaum Associates, 1988.

[6] C. De Souza. *The Semiotic Engineering of Human-Computer Interaction*. MIT Press, 2005.

[7] T. Green. Cognitive dimensions of notations. In *People and Computers V*, pages 443–460. CUP, 1989.

[8] C. A. Gurr. On the isomorphism, or lack of it, of representations. In *Visual language theory*, pages 293–305. Springer, 1998.

[9] R. L. Harris. *Information graphics: A comprehensive illustrated reference*. Oxford University Press, 1996.

[10] M. Jamnik, A. Bundy, I. Green. On automating diagrammatic proofs of arithmetic arguments. *J. of logic, language and information*, 8(3):297–321, 1999.

[11] B. E. John, D. E. Kieras. The goms family of user interface analysis techniques: Comparison and contrast. *ACM TOCHI*, 3(4):320–351, 1996.

[12] P. B. Kohl, N. D. Finkelstein. Student representational competence and self-assessment when solving physics problems. *Physical Review Special Topics-Physics Education Research*, 1(1):010104, 2005.

[13] J. H. Larkin, H. A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1):65–100, 1987.

[14] D. E. Meltzer. Relation between students' problem-solving performance and representational format. *American journal of physics*, 73(5):463–478, 2005.

[15] W. Meulemans, et al. Kelpfusion: A hybrid set visualization technique. *IEEE TVCG*, 19(11):1846–1858, 2013.

[16] D. Moody. The "physics" of notations: toward a scientific basis for constructing visual notations in software engineering. *IEEE TSE*, 35(6):756–779, 2009.

[17] A. Newell. *Unified theories of cognition*. Harvard University Press, 1990.

[18] T. Nipkow, L. Paulson, M. Wenzel. Isabelle/HOL: a proof assistant for higher-order logic. Vol. 2283 *Springer Science & Business Media*, 2002.

[19] D. Raggi, et al. Inspection and selection of representations. In *Int. Conf. on Intelligent Computer Mathematics*, pages 227–242. Springer, 2019.

[20] D. Raggi, et al. Dissecting representations. In *Diagrams*, Springer, 2020. In press.

[21] P. Rodgers, L. Zhang, H. Purchase. Wellformedness properties in Euler diagrams: Which should be used? *IEEE TVCG*, 18(7):1089–1100, 2012.

[22] B. Saket, A. Endert, C. Demiralp. Task-based effectiveness of basic visualizations. *IEEE TVCG*, 25(07):2505–2512, 2019.

[23] A. Shimojima. *Semantic Properties of Diagrams and Their Cognitive Potentials*. CSLI, 2015.

[24] G. Stapleton, M. Jamnik, and A. Shimojima. What makes an effective representation of information: a formal account of observational advantages. *J. of Logic, Language and Information*, 26(2):143–177, 2017.

[25] A. Stockdill et al. Correspondence-based analogies for choosing problem representations in mathematics and computing education. In *IEEE Sym. on Visual Languages and Human-Centric Computing*, 2020.

[26] M. Urbas, M. Jamnik. A framework for heterogeneous reasoning in formal and informal domains. *Diagrams*, pages 277–292. Springer, 2014.

[27] M. W. Van Someren, et al. *Learning with Multiple Representations. Advances in Learning and Instruction Series.* ERIC, 1998.

[28] J. Zhang. The nature of external representations in problem solving. *Cognitive science*, 21(2):179–217, 1997.

[29] J. Zhang, D. A. Norman. A representational analysis of numeration systems. *Cognition*, 57(3):271–295, 1995.